High-Throughput Proteomics: A Flexible and Efficient Pipeline for Protein Production

Sharon A. Doyle, Michael B. Murphy, Jennifer M. Massi, and Paul M. Richardson*

*United States Department of Energy Joint Genome Institute, 2800 Mitchell Drive,
Walnut Creek, California 94598*

Received July 23, 2002

Many studies that aim to characterize the proteome require the production of pure protein in a high-throughput format. We have developed a system for high-throughput subcloning, protein expression and purification that is simple, fast, and inexpensive. We utilized ligation-independent cloning with a custom-designed vector and developed an expression screen to test multiple parameters for optimal protein production in *E. coli*. A 96-well format purification protocol that produced microgram quantities of pure protein was also developed.

Keywords: high-throughput • protein purification • ligation-independent cloning

Introduction

With genome sequencing efforts producing vast amounts of data, attention is now turning toward unraveling the complexities encoded in the genome: the protein products and the cis-regulatory sequences that govern their expression. Understanding the spatial and temporal patterns of protein expression, as well as their functional characteristics on a genomic scale, will foster a better understanding of biological processes from protein pathways to development at a systems level. Several areas of proteomics research are addressing these issues, such as structural genomics studies,[1,2] which aim to characterize the complete repertoire of protein domains in the proteome, and the newly developing field of protein microarray technology,[3] which is attempting to identify protein expression patterns and protein interactions and to catalog them in relational databases. Currently, one of the main bottlenecks in these and many other proteomics initiatives remains the production of sufficient quantities of purified protein.[4] Methods that facilitate protein production in a high-throughput manner are vital to the success of these initiatives.

There are three main steps in the process of protein production: subcloning the protein coding sequence, expression of the soluble protein product in sufficient yield, and purification of the protein from the host proteins. Each of these steps poses unique challenges when applied on a genomic scale. *Eschericia coli* is the most convenient host for high-throughput protein production, although eukaryotic proteins that require post-translational modifications or specific molecular chaperones to promote proper folding are more difficult to produce in bacteria.[5] Affinity fusion partners, such as a hexahistidine peptide or maltose-binding protein, which are attached to the protein coding sequence during the subcloning step, often have dual functions: to promote the production of stable, soluble recombinant proteins in bacteria and to facilitate the parallel purification of multiple samples.[6] Several different tag types and positions (N- or C-terminal) often need to be tested because additional amino acids may also interfere with protein folding, stability, solubility, or function.[7] This increases the need for efficient, high-throughput methods of subcloning and expression analysis. Although several high-throughput systems are available for efficient subcloning, most require an initial time-consuming and carefully planned traditional restriction enzyme based subcloning step. This is not only labor intensive, but also necessitates the screening of all genes of interest for the presence of restriction sites prior to subcloning.

The successful expression of soluble proteins in *E. coli* is dependent on multiple factors, including inducer concentration, induction time, and temperature, as well as host-cell type and growth medium. Identifying conditions for optimal expression of soluble protein is vital because success in purification is often directly related. Once expression conditions have been optimized, high-throughput purification in a 96-well format is possible using relatively small culture volumes.

We describe a system for high-throughput subcloning, protein expression, and purification that is simple, efficient, and flexible. This system differs from others in that customized or expensive robotics are not required,[8] an alternative to recombinatorial cloning methods is used, and an expression screen is used to optimize expression conditions and characterize behavior of the protein during bacterial growth. We utilize ligation-independent subcloning (LIC)[9,10] to create an expression vector encoding hexahistidine-tagged proteins of interest. A dot-blot expression screen is used to analyze total and soluble target protein levels following expression in bacterial cultures, which facilitates the testing of multiple expression parameters. Subsequent protein purification in a 96-well format using immobilized metal affinity chromatography yields highly purified proteins.

* To whom correspondence should be addressed. Phone: (925) 296−5851. Fax: (925) 296−5850. E-mail: pmrichardson@lbl.gov.

10.1021/pr025554a CCC: $22.00 © 2002 American Chemical Society

Published on Web 10/15/2002

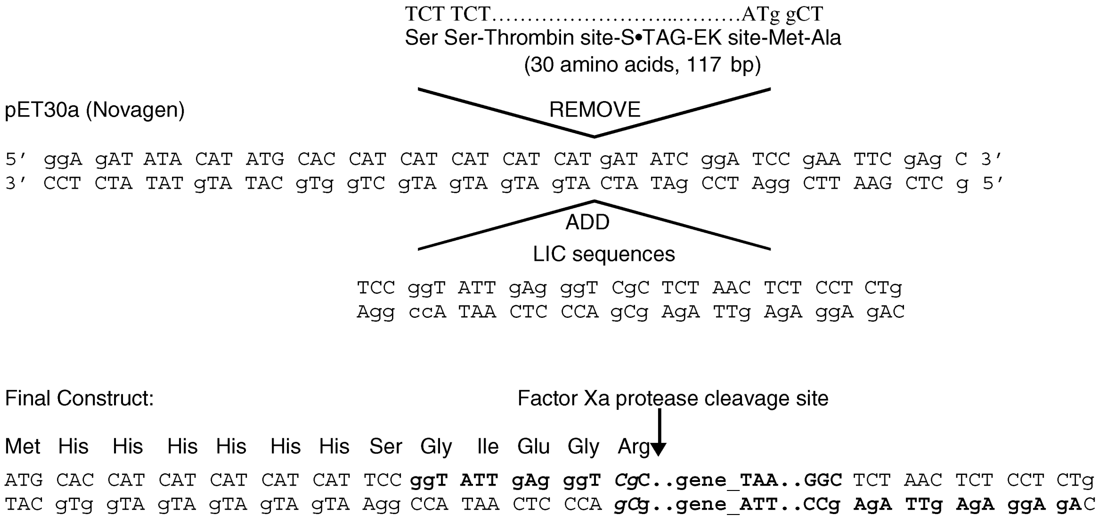

TCT TCT……………………...………ATg gCT
Ser Ser-Thrombin site-S•TAG-EK site-Met-Ala
(30 amino acids, 117 bp)
pET30a (Novagen)
REMOVE
5′ ggA gAT ATA CAT ATG CAC CAT CAT CAT CAT CAT gAT ATC ggA TCC gAA TTC gAg C 3′
3′ CCT CTA TAT gTA TAC gTg gTC gTA gTA gTA gTA CTA TAg CCT Agg CTT AAG CTC g 5′
ADD
LIC sequences
TCC ggT ATT gAg ggT CgC TCT AAC TCT CCT CTg
Agg ccA TAA CTC CCA gCg AgA TTg AgA ggA gAC
Final Construct:
Factor Xa protease cleavage site
Met His His His His His His Ser Gly Ile Glu Gly Arg
ATG CAC CAT CAT CAT CAT CAT TCC ggT ATT gAg ggT CgC..gene_TAA..GGC TCT AAC TCT CCT CTg
TAC gTg gTA gTA gTA gTA gTA Agg CCA TAA CTC CCA gCg..gene_ATT..CCg AgA TTg AgA ggA gAC

Figure 1. Construction of the LIC vector containing the N-terminal hexahistidine affinity tag. The new vector, pNHis, encodes a protein with a N-terminal extension of 6 histidine residues followed by 6 additional amino acids that encode a factor Xa cleavage site. A stop codon was added to the gene sequence so that the 3′ LIC sequence did not add six extra amino acids to the C-terminus of the protein sequence.

Experimental Methods

Subcloning. An expression vector was constructed from pET30 (Novagen) by removing 117 bp 3′ of the hexahistidine tag site that encoded extra affinity tags and by adding the sequence 5′TCCGGTATTGAGGGTCGCTCTAACTCTCCTCTG 3′ to allow for LIC cloning (Figure 1). The new vector, pNHis, was linearized within the LIC sequence by digestion with *Bse*R1, treated with mung-bean nuclease to produce blunt ends, and gel purified. Cloning was performed as described in LIC cloning manuals (Novagen). Briefly, the gene sequences were amplified by PCR using sequence-specific primers with 5′ adaptors (forward primer, 5′ GGTATTGAGGGTCGC 3′; reverse primer, 5′ AGAGGAGAGTTAGAGCCTTA 3′). The insert and vector DNA were treated with T4DNA polymerase in the presence of dGTP and dCTP, respectively, for 40 min at room temperature (22 °C), and the polymerase was heat inactivated for 20 min at 75 °C. The fragments were annealed in a 10 min reaction at room temperature and transformed into NovaBlue competent cells (Novagen). Positive clones were confirmed by a colony PCR procedure using the T7 promoter (5′ TAATACGACTCACTATAGGG 3′) and T7 terminator (5′ GCTAGTTATTGCTCAGCGG 3′) primers, and confirmed by sequencing.

Two sources of coding sequence were used, the bacterium *Xylella fastidiosa*, and *Ciona intestinalis*, a primitive chordate. Eight randomly chosen *Xylella* proteins, ranging from 10 to 32 kDa, were cloned into the expression vector. Five full-length *Ciona* proteins (42 to 88 kDa) were also cloned into this vector, as well as several gene fragments containing DNA binding domains.

Expression Screening. Plasmids encoding *Xylella* genes were transformed into BL21 (DE3) Gold cells (Stratagene) and plasmids encoding *Ciona* genes were transformed into Rosetta pLysS cells (Novagen), based on previous experiments. Initial starter cultures grown at 37 °C were used to inoculate 5 mL LB medium containing 50 μg/mL kanamycin in 24-well blocks. Once an O.D.$_{600}$ of 0.6 to 0.8 was reached, the cultures were induced with IPTG at a concentration of 0.1 mM or 1.0 mM and grown at various temperatures (18, 25, 30, and 37 °C) for 4 h or overnight. In the present study, only temperature, induction strength, and time were tested; however, many other conditions (growth medium, host cells) can be added to the screen. The cells were harvested by centrifugation, frozen at −70 °C, then thawed and resuspended in 0.5 mL lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl, 2 mM $MgCl_2$, 20 mM imidazole, pH 8.0) containing 1 mg/mL lysozyme, 0.5 μL (12U) Benzonase nuclease (Novagen), and 2 μL protease inhibitor cocktail (Sigma). Following incubation on a plate shaker at 4 °C for 30 min, an aliquot of crude lysate was removed, and the remainder of the sample was clarified by centrifugation.

A 2-μL portion of crude and cleared lysate was spotted on Protran nitrocellulose membrane (Schleicher and Schuell) using a 12-channel pipet. A serial dilution (15−1500 ng) of protein standard (isocitrate dehydrogenase, 42 kDa) was also spotted. The membrane was incubated using the Western Processor developing system (Biorad) as follows: TBS (6 mM Tris-Cl, 150 mM NaCl, pH 7.5), 5 min, 3 cycles; blocking buffer (TBS with 3% BSA) 30 min; TBS T/T (TBS with 0.05% Tween 20 and 0.2% Tritin X-100), 5 min, 3 cycles; PentaHis HRP conjugate (Qiagen)-(1:1000 in blocking buffer) 30 min; TBS T/T, 5 min, 5 cycles. The membrane was then treated with metal-enhanced DAB substrate (Pierce), following the manufacturer's protocol, and scanned using a flatbed scanner. The blotting procedure was completed in under 2 h.

Protein Purification. Cleared cell lysates from two wells (10 mL intitial culture) were then batch loaded with 50 μL Ni−NTA Superflow resin (Qiagen) in a Genemate 96-well filter plate and allowed to bind at 4 °C for 20 min with gentle shaking. The column was then formed by applying 200 mbar vacuum pressure. Following three 750 μL washes (50 mM NaH_2PO_4, 300 mM NaCl, 20 mM imidazole, pH 8.0), the protein was eluted with 100 μL wash buffer containing 250 mM imidazole and 20% glycerol followed by 200−400 mbar vacuum pressure.

Determination of Protein Yield and Purity. An Agilent 2100 Bioanalyzer and Protein 200 Plus LabChip kit was used to assess the concentration and purity of the proteins from the 96-well purification. The protein LabChip was prepared by injecting 12 μL of a gel matrix and fluorescent dye mixture into the chip using a chip-priming station. The samples were prepared by mixing 4 μL of protein and 2 μL of a SDS-based denaturing sample buffer containing β-mercaptoethanol, as well as an upper and lower mass standard, and by boiling the mixture. Samples and ladder were then diluted to 90 μL with water and 6 μL of each diluted sample was loaded into a well of the LabChip. Dilution of the samples is necessary to decrease background fluorescence due to the SDS in the sample buffer.[11] The LabChip was then placed in an Agilent 2100 Bioanalyzer, and electrophoresed for 30 min. Agilent Biosizing software was used to determine the size of the proteins of interest by normalization against the two internal standards of 6 and 210 kDa. The fluorescent peak identification settings were adjusted for sensitivity, 0.8 for the minimum peak height, 0.2 s for the minimum peak width, and 4 for the slope threshold. For comparison, 10 μL of each protein sample were run on a traditional 4–20% gradient SDS–PAGE gel (BioRad). The gels were stained with GelCode Blue (Pierce) and scanned on Flour-S MultiImager (BioRad).

Results and Discussion

LIC Cloning. The N-terminal hexahistidine vector used in this study was very efficient for fast and easy subcloning (Figure 1). Because this expression vector encoded an N-terminal tag, the stop codon of the coding sequence was retained in the PCR product insert, so that additional amino acids were not present at the C-terminal end of the expressed proteins. In cases where the stop codon of the target sequence is not utilized (such as when one PCR product is used for multiple constructs with both N- and C-terminal affinity tags), only six additional amino acids are added to the C-terminal end of the protein using the LIC system.

One major advantage of the LIC cloning system is its flexibility; any vector can be made into a LIC vector simply by inserting the LIC cloning sequence (Figure 1). Generation and isolation of the linear expression vector used in the annealing requires restriction digestion, gel separation, and purification; however, this can be performed on a large scale to provide the annealing vector for many reactions. The annealing reactions are fast and efficient and do not require expensive enzymes. The percentage of clones containing the expected insert is generally greater than 90%, ensuring that only a few colonies need to be tested by PCR methods to identify positive clones. The high success rate and ease of each step makes this process scalable. Also, these methods provide the opportunity for cloning into several expression vectors so that multiple affinity tags can be tested for each target gene, increasing the success of finding conditions that promote high expression levels. This study used an N-terminal hexahistidine peptide tag; however, additional vectors that incorporate a C-terminal hexahistidine peptide and N-terminal maltose binding protein have been generated as well.

Expression Screening. Expression constructs from *Xylella fastidiosa* and *Ciona intestinalis* were tested under various growth conditions using the dot blot procedure to identify optimal growth conditions for each protein (Figure 2). The standard curve generated on each blot was reproducible and was useful to approximate the concentrations of protein in the sample spots. A good correlation between spot intensities and protein concentration was obtained.

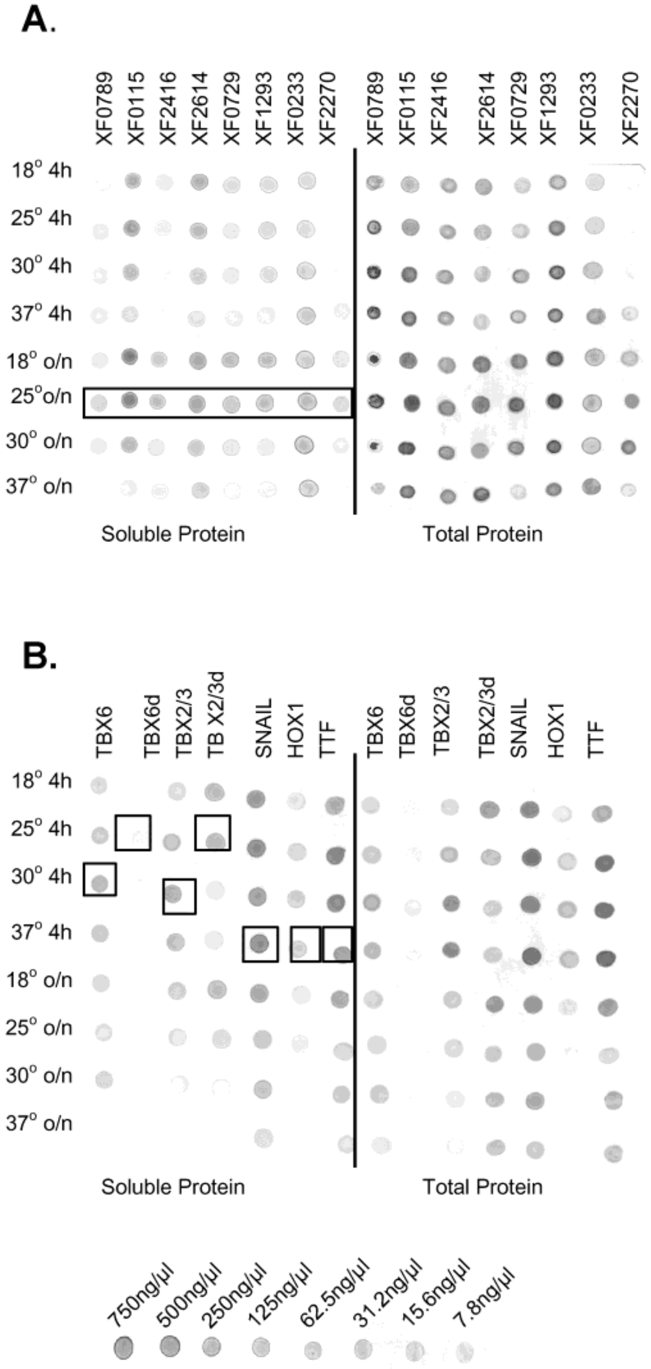


Figure 2. Expression screen dot blots of (A) *Xylella* samples and (B) *Ciona* samples, grown for 4 h or overnight (o/n). Below is shown the standard curve. The spots outlined with squares indicate the samples chosen for further purification.

Overall, the *Xylella fastidiosa* proteins showed good total expression levels, with nearly all of the crude lysate samples exhibiting high-intensity spots (Figure 2A). Two samples (proteins XF0233 and XF2614) showed little difference in the total protein and soluble protein samples under all conditions tested, indicating that these proteins are very soluble. For the remainder of the samples, however, a trend was seen where the solubility was increased with reduced growth temperature (≤25 °C). In addition, overnight induction conditions produced better protein yields than those found in 4 h induction conditions. Nearly identical results were obtained when samples were induced with 0.1 and 1.0 mM IPTG (data not shown).

The total expression levels for the *Ciona intestinalis* samples were not as consistent as in the *Xylella* set (Figure 2B). Unlike the *Xylella* proteins, total expression was better in samples

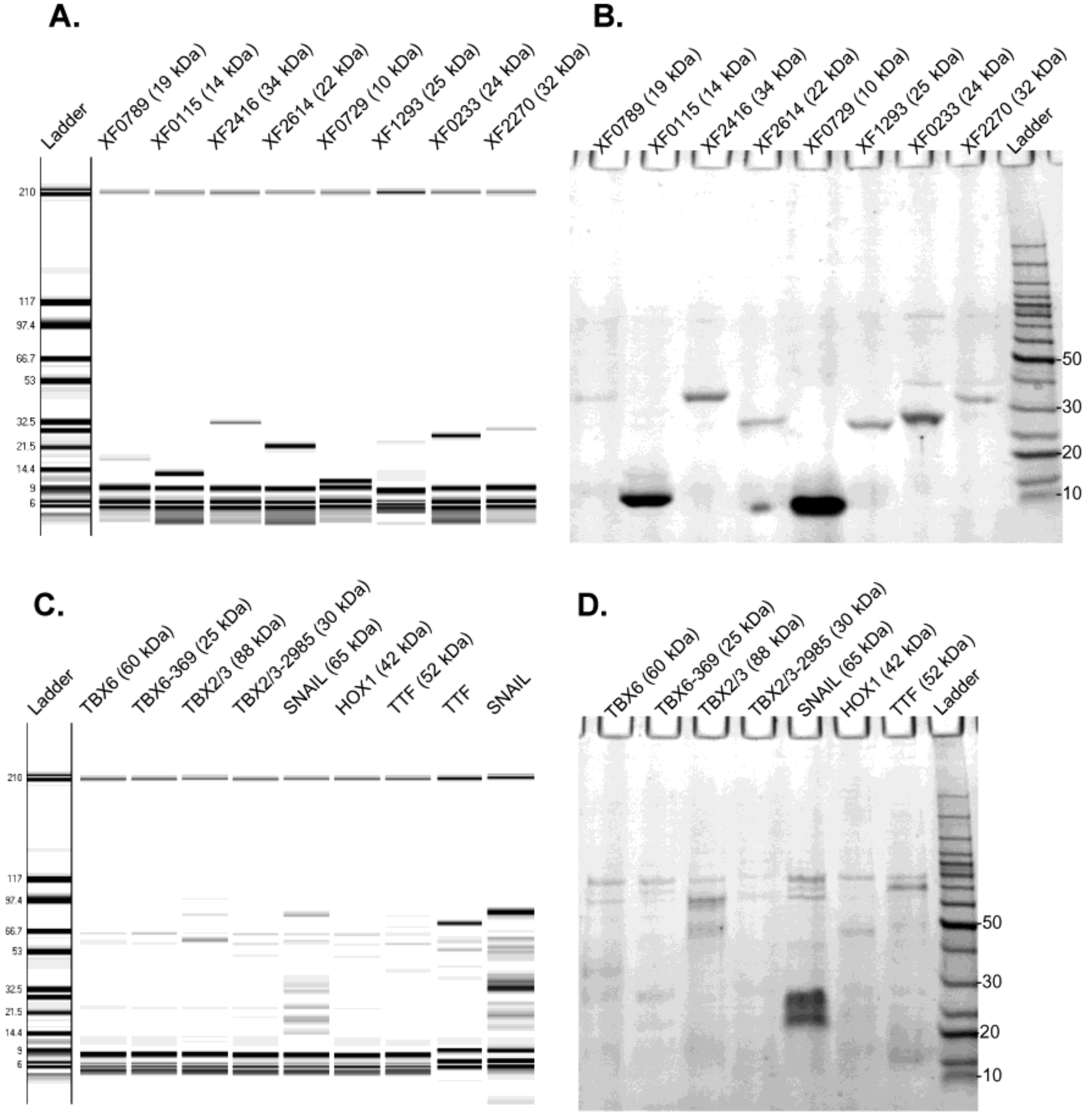
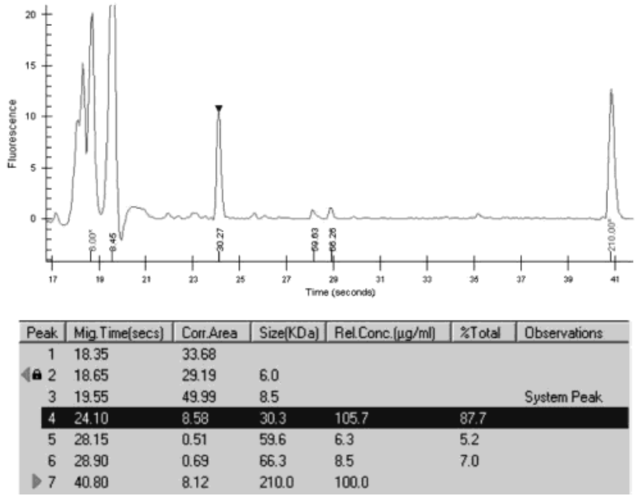


Figure 3. Purification results of *Xylella* (A, B) and *Ciona* (C, D) protein samples run on the Agilent Bioanlyzer (A, C) and SDS−PAGE (B, D). Four-μL samples were run on the Bioanalyzer, and 10-μL samples were run on the SDS−PAGE gels. In the Bioanalyzer samples, the internal upper and lower mass standards (6 and 210 kDa) are added to the purified protein when the sample is loaded onto the Agilent Chip. The last 2 lanes of panel C show results from the scaled up purification of snail and TTF-1 proteins.

induced for 4 h than those of samples induced overnight, although not at every temperature tested. Snail and TTF-1 expressed soluble protein under most conditions, whereas the remainder of the proteins showed reduced solubility when induced overnight. In the 4 h samples, snail and TTF-1 produced their highest soluble yields at ≥25 °C, where 25, 30, and 37 °C were roughly equivalent. For the proteins showing lower expression levels, no clear pattern was seen. Hox1 produced more soluble protein at ≥25 °C, with the highest yield at 37 °C, whereas Tbx2/3 and Tbx6 produced the highest yields of soluble protein at 30 °C. These differences are probably due to the inherent stability of these proteins when expressed in bacteria. As with the *Xylella* samples, nearly identical results were obtained when the same experiments were performed with 0.1 and 1.0 mM IPTG induction (data not shown).

Significant differences were seen in the expression of full-length and DNA binding domains of *Ciona* proteins Tbx6 and Tbx2/3 (Figure 2B). The domain of Tbx6 showed little or no expression under any conditions, unlike the full-length protein, which expressed fairly well at higher temperature for 4 h. The TBX2/3 domain (Tbx2/3d) clearly showed highest yields at ≤25 °C, whereas the full-length protein expressed more soluble protein at ≥25 °C. These results suggested that the cloned DNA fragments do not contain structurally stable DNA binding domains, and thus, the smaller protein products were less stable than the full-length proteins. Comparison of additional full-length proteins and fragments will provide useful information regarding the ability of protein fragments containing specific domain motifs to fold into stable protein products when expressed in bacteria.

This expression screen allows for the identification of optimal conditions for soluble protein expression in a convenient and reproducible manner. Other methods for expression screening using microarrays have been described;[12] however, our method focuses on the use of a screen to determine the amount of total and soluble protein produced under a variety of growth



Peak	Mig.Time(secs)	Corr.Area	Size(KDa)	Rel.Conc.(µg/ml)	%Total	Observations
1	18.35	33.68				
2	18.65	29.19	6.0			
3	19.55	49.99	8.5			System Peak
4	24.10	8.58	30.3	105.7	87.7	
5	28.15	0.51	59.6	6.3	5.2	
6	28.90	0.69	66.3	8.5	7.0	
7	40.80	8.12	210.0	100.0		

Figure 4. Analysis of protein purity and yield using the Agilent Bioanlayzer. This figure shows the results of the Xylella protein XF2270 (Figures 2 and 3, panel A). The electropherogram identifying protein peaks is shown, with the upper and lower mass standards marked as 6 and 210 kDa, respectively. The 9 kDa System Peak is also identified. The remaining peaks, 30 kDa protein XF2270 and contaminant proteins (59 and 66 kDa) are also marked. The purity level of the XF2270 sample is 87.7% as judged by the Biosizing software.

parameters to identify optimal conditions for protein expression. This is critical to the success of a protein production process, even when all clones are sequence verified and in the correct reading frame, because success in protein purification is directly related to the expression level of the protein. In addition, the use of 24-well blocks in standard incubators for cell growth, standard lysis procedures, and simple steps of centrifugation and sample spotting on nitrocellulose make it easily implemented, without the need for expensive robotics. In many cases, protein expression levels are sufficiently high to allow for direct purification of micrograms of protein from 5 or 10 mL cultures grown in the 24-well blocks. Alternatively, if more protein is required, the optimal conditions for scaled up experiments can be identified using this method.

Protein Purification. Purification of the sample proteins from the cultures grown in 24-well blocks was performed in 96-well filter plates using a standard vacuum manifold for the column forming, washing, and eluting steps. Protein samples were run on an Agilent 2100 Bioanalyzer using a Protein 200 Plus LabChip for protein purity, yield, and concentration determinations (Figure 3, panels A and C). This system offers the advantages of requiring only 4 μL of protein samples, as compared to the 10 μL required for equivalent band identification on traditional SDS−PAGE gels (Figure 3, panels B and D) and produces results in 30 min, compared to several hours for SDS−PAGE. The Biosizing software provides detailed tables of raw and analyzed data and gel images and is useful for data storage and retrieval (Figure 4). The Agilent system provides accurate protein concentrations, alleviating the need of running additional assays or using more of the sample for characterization. In addition, because it determines the concentration of individual proteins or "bands" on the gel image, accurate concentrations of partially pure proteins can be easily obtained.

Protein yields for the *Xylella* samples ranged from 10 μg to 100 μg, which correlated well with predictions from the expression screen (Figure 3A). The finding that the total protein quantities did not decrease after prolonged growth (unlike the *Ciona* samples), and that the soluble protein spots increased in intensity with prolonged growth suggested that the proteins were stable, was predictive of successful purification. Only one sample (XF1293) did not purify as expected, probably due to degradation during the purification procedure. The purity of the *Xylella* proteins was high, ranging from 92 to 100%, as judged by the Bioanalyzer software. From a test set of over 40 randomly chosen *Xylella* proteins, over 70% were successfully purified (data not shown).

Overall, the *Ciona* samples produced similar amounts of soluble proteins as the *Xylella* samples when induced for 4 h at a variety of temperatures; however, purification was not as successful (Figure 3C). Nearly all of the samples produced less protein as the induction time and temperature were increased, suggesting that protein stability was a potential problem. This was not surprising, as these proteins have a higher molecular mass than the *Xylella* set and are eukaryotic, which increases the possibility that protein folding or stability of the folded structure is affected by being generated in a non-native environment. Although the 4 h samples showed optimal soluble protein levels in the expression screen, much of the protein may have been in an partially unfolded state but not yet in the form of insoluble inclusion bodies, resulting in an overestimation of soluble protein.[7] Additionally, some of the protein may have been partially degraded, yet still bound by the nitrocellulose membrane, which would also contribute to a difference in the amount of predicted and produced full-length protein. Western blots and mass spectrometry of many of the major bands of both the snail and TTF-1 samples confirmed that they were degradation products and not contaminating proteins. The samples grown at lower temperatures that exhibited similar protein levels in the expression screen may have been better candidates for purification because they may contain less degraded protein, which would have reduced the discrepancies between the expression screen and purification results.

The purity of most of the *Ciona* protein samples was only up to 30%, and positive band identification was sometimes difficult. This is common when expression levels are low and can be improved by scaling up the purification process, as was done for the snail and TTF-1 samples. Purifications from 1-L cultures of these proteins using the growth conditions identified in the screen that optimized protein-to-column ratios improved the purity and yield of the purified products (Figure 3). Additionally, affinity fusions such as a C-terminal hexahistidine or maltose binding protein that may improve expression can be easily tested using the LIC cloning and expression screen protocols, as well as a broader range of expression conditions.

Conclusions

Our results show that LIC cloning is an ideal high-throughput cloning method that is easy, reliable, and flexible. The dot blot expression screen is a convenient way to test multiple parameters for optimal protein expression, or to identify the response of recombinant proteins to different growth conditions. Once proteins are segregated based on the results of the screen, they can proceed to the purification process. In some cases, such as for the *Xylella* proteins, the expression screen identified that one growth condition appeared to be acceptable for all samples, thus possibly alleviating the need for future screening of individual proteins. More challenging examples, such as the *Ciona* proteins that were not easily expressed in soluble form, can be subjected to additional screens or processed in a larger scale format to produce adequate protein for downstream use.

As the number of proteins that are tested in the expression screen increases, analysis of the results may provide valuable insight into the relationship between types of domains and their soluble expression in bacteria. This analysis will also provide information on the ability of specific affinity fusions to enhance the expression of soluble protein as well as provide optimal yield and purity of purified proteins. This system provides an efficient and effective work flow for current high-throughput protein production and will produce data that will aid in the development of predictive methods to further improve this process.

Acknowledgment. We thank Shirin Fuller and Joann Wang for technical assistance and Dr. Peter Beernink for his critical reading of the manuscript. This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory (Grant No. W-7405-Eng-48), Lawrence Berkeley National Laboratory (Grant No. DE-AC03-76SF00098), and Los Alamos National Laboratory (Grant No. W-7405-ENG-36).

References

(1) Burley, S. K. *Nat. Struct. Biol.* **2000**, *7*, 932−934.
(2) Terwilliger, T. C. *Nat. Struct. Biol.* **2000**, *7*, 935−939.
(3) Walter, G.; Bussow. K.; Lueking, A.; Glokler, J. *Trends Mol. Med.* **2002**, *8*, 250−253.
(4) Bouguslavsky, J. *Genomics Proteomics* **2001**, *1*, 44−46.
(5) Makrides, S. C. *Microbiol. Rev.* **1996**, *60*, 512−538.
(6) Nilsson, J.; Stahl, S.; Lundeberg, J.; Uhlen, M.; Nygren, P. *Protein Expression Purif.* **1997**, *11*, 1−16.
(7) Braun, P.; Hu, Y.; Shen, B.; Halleck, A.; Koundinya, M.; Harlow, E.; LaBauer, J. *Proc. Natl. Acad. Sci., U.S.A.* **2002**, *99*, 2654−2659.
(8) Lesley, S. A. *Protein Expression Purif.* **2001**, *22*, 159−164.
(9) Aslanidis, C.; De Jong, P. J. *Nucleic Acids Res.* **1990**, *18*, 6069−6074.
(10) Haun, R. S.; Servanti, I. M.; Moss, J. *Biotechniques* **1992**, *13*, 515−518.
(11) Bousse, L.; Mouradian, S.; Minalla, A.; Yee, H.; Williams, K.; Dubrow, R. *Anal. Chem.* **2001**, *73*, 1207−1212.
(12) Lueking, A.; Horn, M.; Eickhoff, H.; Bussow, K.; Lehrach, H.; Walter, G. *Anal. Biochem.* **1999**, *270*, 103−111.

PR025554A